ISSN:0975-9646

T.Sumallika et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (1) , 2014, 617 - 620

Automatic User Specific Opinion Extraction from Online Reviews

T.Sumallika[1], P.Pranathi[2] ,Sk.Ramzan Bibi[2] ,P.Aparna[2] , D.Shanti Lakshmi [2] ,R.GowtamKumar[2]

[1]Asstistant Professor, [2]B.Tech student
Dept. of CSE *LIET,*
Vizianagaram,India

Abstract-Typically for buying any kind of goods or to know the survey about any particular product we generally look for message boards, web content, blogs, news to know reviews. Generally reviews are based upon the sentiment identification phase, which associates expressed opinions with each relevant entity and scoring techniques. Our system uses natural language processing techniques to assist the customer in buying products based upon the online reviews. We enhanced nearest-adjective algorithm that uses parser and tagger to produce a report of a product and have also used PMI (point wise mutual information) algorithm for comparison purpose and gather all relative nouns. Our project is a step ahead which has three analysis levels namely overall report, sentence level analysis, review level analysis. These analysis processes helps in explaining about the products individual feature in brief.

Keywords: PMI algorithm, Nearest Adjective Algorithm, Pos Tagger, Penntree Bank Tag Set.

1. INTRODUCTION

Basically reviews can be positive, neutral or negative. But this doesn't give brief expression of each and every feature of the product. Previously sentiment analysis was done based on newspapers and blogs review. Opinions of news entities about people, places and things for which the system assigned scores indicating positive or negative view of each distinct entity in the text corpus. Sentiment identification phase, associates expressed opinions with each relevant entity, and a sentiment aggregation and scoring phase, which scores each entity relative to others in the same class. Sentiment analysis aims to determine the attitude of a speaker or a writer with respect to some topic or the overall tonality of a document. In recent years, the exponential increase in the Internet usage and exchange of public opinion is the driving force behind Sentiment Analysis today. The Web is a huge repository of structured and unstructured data. The analysis of this data to extract latent public opinion and sentiment is a challenging task.

Word of mouth (WOM) is the process of conveying information from person to person and plays a major role in customer buying decisions. In commercial situations, WOM involves consumers sharing attitudes, opinions, or reactions about businesses, products, or services with other people. People depend on families, friends, and others in their social network. Research also indicates that people appear to trust seemingly disinterested opinions from people outside their immediate social network, such as online reviews. This is where Sentiment Analysis comes into play. Growing availability of opinion rich resources like online review sites, blogs, social networking sites have made this "decision-making process" easier for us. With explosion of Web 2.0 platforms consumers have a soapbox of unprecedented reach and power by which they can share opinions. Major companies have realized these consumer voices affect shaping voices of other consumers.

2. TECHNIQUES OF OPINION EXTRACTION

We have applied different techniques using various algorithms to extract opinions for the online reviews which are briefly mentioned below

2.1. Point wise mutual information (PMI)

It is information theory approach to find collocation.
Collocation is an expression of two or more words that are some predictable way of saying something. In simple words it is measure of how much every single word tells about the other word.

NUMERICAL REPRESENTATION: Let us consider two words l and m in the given review a particular product then,

Formulaes:

$$I(l, m) = \log_2 \frac{P(l, m)}{P(l)\,P(m)}$$

$$= \log_2 \frac{P(l|m)}{P(l)}$$

$$= \log_2 \frac{P(m|l)}{P(m)}$$

Bigram frequency: it is every sequence of two adjacent elements in a string of tokens which are typically letters, symbols or words. Suppose we are taking sample comments of a product to extract an opinion of customers we need to compare adjacent words of a comment to get accurate result.

Formulae:

$$P(u_n | u_{n-1}) = \frac{P(u_{n-1}, u_n)}{P(u_{n-1})}$$

The PMI is used for two different tasks:

(i) To find the adjacent word that occur together most frequently
(ii) To generate pairs between long distance words

Long distance PMI:

Formulae:

$$I_d(l, m) = \log_2 \frac{P_d(l, m)}{P_d(l)\,P_d(l)}$$

In order to get review for more than two words we extend the PMI algorithm to formulae with three words.

2.2. Nearest-adjective algorithm

Considering our review page or customer opinion similar to travel sales man problem we have obtained the following algorithm:

(i) Assume a word has arbitrary vertex or current vertex V.
(ii) Using PMI algorithm find the nearest similar word and connect to the word.
(iii) Set the current word has V.
(iv) Mark V has visited.
(v) If all the words are visited, then terminate.
(vi) Go to step 2.

This algorithm states that the nearest adjective to a feature speaks about the feature. For example, consider the following review:

I was very happy with the product. It looks brand new and plus everything came in the box as promised. Fast delivery as well. Love it!

Here, *happy* is the adjective that is nearest to the noun *product.* The process continues for the whole document (set of reviews). This works most of the times if we have the database of orientations of adjectives.

Limitations:
(i) Having a database that handles all the adjectives is impractical.
(ii) This algorithm is a document level analysis algorithm, hence if a noun of one sentence is Nearer to the adjective of another sentence, that is considered instead of the same sentence adjective. This is a serious issue.
(iii) The nearest adjective to a feature need not speak about the feature all the time.

We address solutions to both these limitations.
The unidentified adjectives are made a list along with the sentences where they appear. This list is done based on the frequency of appearance of adjectives in the document. The user can categorize adjectives whenever he/she wants to, which results in the ever-increasing database.
The second problem can be solved by making it a sentence level analysis algorithm.
The third problem is solved by deploying a standard parser. The parser gives exact adjective which is dependent on the particular feature.

2.3. Alphabetical list of part-of-speech tags used in the Penn Treebank Project:

POS TAGGER
The process of classifying words into their parts of speech and labelling them according is known as parts of speech tagging
The collection of tags used for a particular task is known as a tag set.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predetermined
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Example:
"WELL, ITS MAIN PROBLEM IS THAT IT'S SIMPLY TOO JUMBLED"
RB well, Prb$ its JJmain NNproblem VBZ is IN PRP it VBZ'S RB simply RB too jumbled
RB----------------adverb
PRP$--------------possessive pronoun
JJ-------------------adjective
NN-----------------noun
VBZ---------------verb, 3rd person singular person
IN------------------preposition
PRP------------------personal pronoun

- *Ever term has been associated with a relevant log indicating its role in the sentence such as VBZ (verb) NN (NOUN)*
- *The entire list of tag & their meaning is based on the Penn tree bank tag set*
- *Then we take the adjective in the words*
- *Now t can identify the frequent and infrequent features*

3. LEVELS OF OPINION ANALYSIS

As mentioned above we have three levels of opinion analysis to present clear view of a product to customer.

3.1.1 Review level analysis

Each review is assigned a review coefficient, whose value varies between -1.0 and +1.0, the negative coefficient indicates the review speaks negative about the product and positive coefficient indicates the review speaks positive. +1.0 is for most positive, -1.0 for most negative and 0.0 for neutral sentiment. This is done by using conventional nearest-adjective algorithm rather than proposed enhanced algorithm.

3.2.2 Sentence level analysis

The features selected by the user are taken as final features and the acquired user reviews are processed sentence by sentence to look for the final features and if the sentence is believed to be speaking about a particular feature of the product it is considered an opinion sentence and added to the sentence level analysis report.

3.2.3 Overall report analysis

As in the sentence level analysis, sentences are looked for the features. In this analysis, the whole document is searched for the features and this gives only the percentage of opinion sentences that speak positive about a particular feature, this is done for all the user-selected features.

4. WORKING:

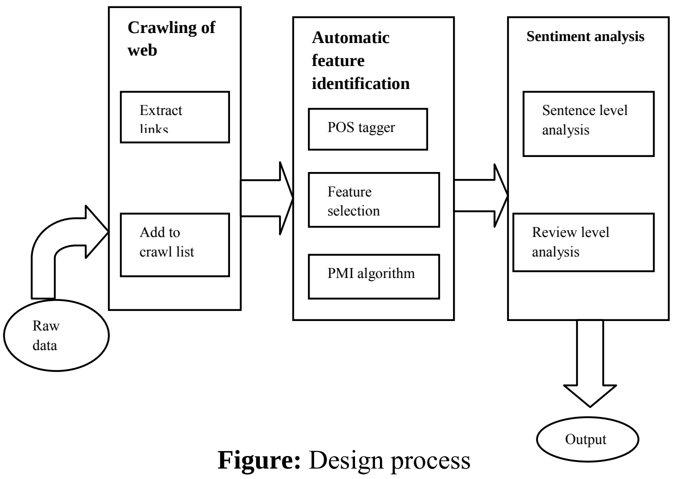

Figure: Design process

4.1. Crawling of web

The product information from the user is taken and searched for user specific data from Google and extract the links are extracted from the Google page.

These links are processed and the links are added to the list of links to be crawled. These links from the list are processed to get relevant data and links from this web page are added to the list if they are not already crawled. This process continues until the user specified numbers of reviews are extracted from the web.

4.2. Automatic feature identification

After retrieving the user opinions from the web, POS tagger tags the sentences.

Two types of features are identified in this process. They are: (i) Frequent features

Frequent features are the nouns that appear the most number of times; these nouns must come after article 'the'. Top n features are identified as frequent features

(ii)Infrequent features.

The infrequent features are the ones that appear as nouns with the same orientation as that of top frequent nouns.

4.3 Sentiment analysis

The identified features are displayed to the user so that the user can select features according to his/her wish. Each sentence in each review is searched for one of the user selected features and the adjective which speaks about the feature form a pair, which are used to polarize the sentence. This process continues for all the reviews.

According to the (noun, adjective) pairs feature wise analysis is displayed as an output. The process is done in two levels: Sentence Level, Review Level.

Sentence Level Analysis yields sentences that speak about each feature where they are categorized feature wise.

Example: here we take an example of a mobile XYZ and review of each feature is mentioned as below:

- *DEVICE: The XYZ mobile is a high performance device that is recommended to everyone……>>we were more excited about the 8mp camera than actually taking pictures in the store otherwise we most likely would have chosen a different device.*
- *DESIGN: The XYZ mobile is big in size hence it is not feasible to carry…>>hence this is the major drawback.*
- *BATTERY: very durable, great battery life, great call quality…>>this is major reason why customers are attracted towards this product.*

Review Level Analysis assigns a review coefficient to each review which indicates the usefulness of the review.

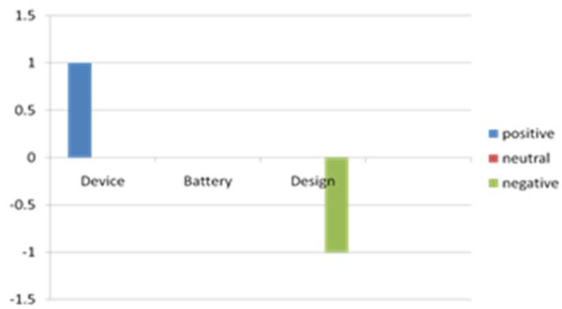

Figure: Review analysis of mobile XYZ

In the above graph we have clearly shown review level analysis of mobile XYZ. Each feature of the mobile is shown as positive negative or neutral based upon the customers feedback comments. The device, battery, design are shown has positive, neutral and negative respectively.

5. CONCLUSION:

Opinions are a unique type of information that is different from facts. The methods for content classification based on ranking (like those used by search engines) are not effective or simply do not accurately depict reality, as one opinion is different from multiple opinions.

It is feasible and reliable to build system capable of classifying and organizing opinions through the so-called feature-based summary, which resumes the most relevant information for users. However, it is undeniable that a great number of opinions are difficult to classify due to the complexity of the human language.

While seeking for a review a customer generally watch the top most comments in the comment session and comes to a conclusion, but there are possibility of viewing more number of negative comments in the front page of comment session. With our project we scan each and every comment and make customer receive the accurate review about the product.

Evaluation also showed that the system can be more effective when domain specific, using the help of manual annotations to treat common exceptions. A system can therefore combine multiple approaches with the intelligence of automatic algorithms and manual annotations in order to provide a high degree of accuracy.

This is what we address in this project by enhancing the existing nearest-adjective algorithm that comparatively better results than the original one and even PMI (point wise mutual information) deploying algorithm.

The work can be further extended to emerging areas to investigate with soft computing techniques like neural network.

REFERENCES:

[1] Zhu, Jingbo Wang, Huizhen Zhu, Muhua Tsou,Benjamin K. Ma, Matthew, “Aspect-Based Opinion Polling from Customer Reviews”, IEEE Transactions on Affective Computing, Volume: 2,Issue:1 On page(s): 37. Jan-June 2011.

[2] Yi, J., T. Nasukawa, R. Bunescu, and W. Niblack: 2003, “Sentiment Analyzer: Extracting Sentiments about a Given Topic using Natural Language Processing Techniques”, In: Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM-2003). Melbourne, Florida.

[3] Alekh Agarwal and Pushpak Bhattacharyya, “Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified”, In Proceedings of the International Conference on Natural Language Processing (ICON), 2005.

[4] Ahmed Abbasi, Hsinchun Chen, And Arab Salem, “Sentiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums”, ACM Trans. Inf. Syst., Vol. 26, No. 3. (June 2008), pp. 1-34.

[5] Anindya Ghose, Panagiotis G. Ipeirotis, “Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews”, Proceedings of the Ninth International conference on Electronic commerce ICEC07 (2007), pp: 303-310.

[6] Bing Liu,"Exploring User Opinions in Recommender Systems", Proceeding of the second KDD workshop on Large Scale Recommender Systems and the Netflix Prize Competition", Aug 24, 2008, Las Vegas, Nevada, USA.

[7] Dave.D, Lawrence.A, Pennock.D,"Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews", Proceedings of International World Wide Web Conference (WWW‟03), 2003.

[8] Turney, P, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews", ACL‟02, 2002.

[9] Lina Zhou, Pimwadee Chaovalit**,** “Movie Review Mining: a Comparison between Supervised and Unsupervised Classification Approaches”, Proceedings of the 38th Hawaii International Conference on system sciences, 2005.

[10] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, “Thumbs up? Sentiment classification using machine Learning techniques”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 79–86, 2002.